1. What is polarity detection?

= Sentiment polarity for an element defines the orientation of the expressed sentiment, i.e., it determines if the text expresses the positive, negative or neutral sentiment of the user about the entity in consideration. Polarity detection is the most common and essential task of sentiment analysis. Polarity detection is a binary classification task that represents an important dowel in most sentiment analysis applications. It is a process of identifying and categorizing opinions expressed in a piece of text. It classifies the text into positive, negative or neutral.

2. What are the drawbacks of RNN? how can this be overcomed? What is the advantage of using Attention mechanism?

= The drawbacks of RNNs are given below:

1. Training RNNs is different
2. The vanishing or exploding gradient problem
3. RNNs cannot be stacked up
4. Slow and Complex training procedures
5. Difficult to process longer sequences

The problem of exploding gradients can be solved using gradient clipping. As the name suggests, the gradients are clipped once they reach a pre-defined threshold. But the problem of vanishing gradients is still out there. It was later solved up to a point with the introduction of LSTM networks.

It is quite difficult to train RNNs on too long sequences, especially while using ReLU or tanh activations. This is another reason for introducing the GRU based networks.

RNNs cannot remember longer sentences and sequences due to the vanishing/exploding gradient problem. It can remember the parts which it has just seen. Although an LSTM is supposed to capture the long-range dependency better than the RNN, it tends to become forgetful in specific cases. Due to this, the “Attention” mechanism was suggested that not only can all the input words be taken into account in the context vector, but relative importance should also be given to each one of them. The advantages of attention is its ability to identify the information in an input most pertinent to accomplishing a task, increasing performance especially in natural language processing - Google Translate is a bidirectional encoder-decoder RNN with attention mechanisms.

3. What was the evaluation criteria used for comparing the models?

= Four evaluation criteria have been used to compare the models. They are: Precision (Pr), Recall (Re), F1-measure (F1), and Accuracy (Acc).

4. What gates are used in LSTM? What are their features?

= A common LSTM unit is composed of a cell, an input gate, an output gate and a forget gate. The cell remembers values over arbitrary time intervals and the three gates regulate the flow of information into and out of the cell.

The forget gate decides what information should be thrown away or kept. Information from the previous hidden state and information from the current input is passed through the sigmoid function. Values come out between 0 and 1. The closer to 0 means to forget, and the closer to 1 means to keep.

To update the cell state, we have the input gate. First, we pass the previous hidden state and current input into a sigmoid function. That decides which values will be updated by transforming the values to be between 0 and 1. 0 means not important, and 1 means important. You also pass the hidden state and current input into the tanh function to squish values between -1 and 1 to help regulate the network. Then you multiply the tanh output with the sigmoid output. The sigmoid output will decide which information is important to keep from the tanh output.

Last, we have the output gate. The output gate decides what the next hidden state should be. The hidden state contains information on previous inputs. The hidden state is also used for predictions. First, we pass the previous hidden state and the current input into a sigmoid function. Then we pass the newly modified cell state to the tanh function. We multiply the tanh output with the sigmoid output to decide what information the hidden state should carry. The output is the hidden state. The new cell state and the new hidden is then carried over to the next time step.

Basically, the forget gate decides what is relevant to keep from prior steps. The input gate decides what information is relevant to add from the current step. The output gate determines what the next hidden state should be.

5. Why performance of ABCDM was better on review datasets than twitter dataset?

= ABCDM outperformed the other six models in terms of both accuracy and F1 measures with Twitter datasets. However, the amount of improvements is less than using the review datasets. The key reason for this is Twitter datasets contains small number of words. ABCDM does not yield significant improvement using short comments, because the first feature extraction layer in this model is an RNN-based network which is designed to capture long dependencies. The rationale behind the higher improvement for positive reviews as compared to the negative ones does not hold, because every tweet contains about 15 words in average which may be more than one sentences. This, results in having very short sentences in which there is not long distances between negation and other sentiment polarity changers. Therefore, there is no much differences between the structure of positive and negative tweets.